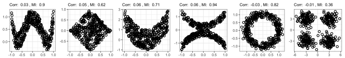
Introduction to Machine Learning

Mutual Information under Reparametrization (Deep-Dive)



Learning goals

- Understand why MI is invariant under certain reparametrizations

MUTUAL INFORMATION PROPERTIES

- MI is invariant w.r.t. injective reparametrizations that are in $\mathcal{C}^1$:

  Let $f, g : \mathbb{R}^d \to \mathbb{R}^d \in \mathcal{C}^1$ be injective transformations and X, Y be continuous random variables in $\mathbb{R}^d$ then by the change of variables the joint and marginal densities of $\tilde{X} = f(X)$, $\tilde{Y} = g(Y)$

$$\begin{aligned} \tilde{p}(\tilde{x}, \tilde{y}) &= p(f^{-1}(\tilde{x}), g^{-1}(\tilde{y})) \cdot |J_{f^{-1}}(\tilde{x})| \cdot |J_{g^{-1}}(\tilde{y})|, \\ \tilde{p}(\tilde{x}) &= p(f^{-1}(\tilde{x})) \cdot |J_{f^{-1}}(\tilde{x})|, \quad \tilde{p}(\tilde{y}) = p(g^{-1}(\tilde{y})) \cdot |J_{g^{-1}}(\tilde{y})|, \end{aligned}$$

  where $p(x, y)$ is the joint density of X and Y and $p(x), p(y)$ are the respective marginal densities. (J denotes the Jacobian)

  With this, it follows that

$$I(\tilde{X}; \tilde{Y}) = \int \tilde{p}(\tilde{x}, \tilde{y}) \log \left(\frac{\tilde{p}(\tilde{x}, \tilde{y})}{\tilde{p}(\tilde{x}) \tilde{p}(\tilde{y})} \right) d\tilde{x} d\tilde{y} = *$$

MUTUAL INFORMATION PROPERTIES

$$
\begin{aligned}
* = &\int p(f^{-1}(\tilde{x}), g^{-1}(\tilde{y})) \cdot |J_{f^{-1}}(\tilde{x})| \cdot |J_{g^{-1}}(\tilde{y})| \\
&\cdot \log\left(\frac{p(f^{-1}(\tilde{x}), g^{-1}(\tilde{y})) \cdot |J_{f^{-1}}(\tilde{x})| \cdot |J_{g^{-1}}(\tilde{y})|}{p(f^{-1}(\tilde{x}))|J_{f^{-1}}(\tilde{x})| \cdot p(g^{-1}(\tilde{y}))|J_{g^{-1}}(\tilde{y})|}\right) d\tilde{x}d\tilde{y} \\
= &\int p(f^{-1}(f(x)), g^{-1}(g(y))) \cdot |J_{f^{-1}}(f(x))| \cdot |J_{g^{-1}}(g(y))| \\
&\cdot \log\left(\frac{p(f^{-1}(f(x)), g^{-1}(g(y)))}{p(f^{-1}(f(x)))p(g^{-1}(g(y)))}\right) |J_f(x)| \cdot |J_g(y)| dxdy \\
= &\int p(x, y) \cdot |J_{f^{-1}}(f(x))J_f(x)| \cdot |J_{g^{-1}}(g(y))J_g(y)| \log\left(\frac{p(x, y)}{p(x)p(y)}\right) dxdy \\
= &\int p(x, y) \cdot \log\left(\frac{p(x, y)}{p(x)p(y)}\right) dxdy = I(X; Y).
\end{aligned}
$$

(The fourth equality holds by the inverse function theorem)